Creation of Time-Varying Voiceprint Database

Linlin Wang and Thomas Fang Zheng
Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
wangll07@mails.tsinghua.edu.cn, fzheng@tsinghua.edu.cn

Abstract

Performance degradation with time varying in speaker recognition is a generally acknowledged phenomenon and it is widely assumed that speaker models should be updated from time to time to maintain representativeness. However, lack of a longitudinal voiceprint database which specially focuses on the time-varying effect has prevented researchers from finding out reasons behind this phenomenon. In this paper, an on-going voiceprint database creation project is presented, aiming to examine solely the time-varying impact on speaker recognition and explore hidden factors causing possible performance degradation. In this speech database, speakers are requested to utter in a reading way with fixed prompt texts instead of free-style conversations throughout 16 sessions in a period of about three years. Sessions are of gradient time intervals where initial ones are of shorter time intervals and following ones of longer and longer time intervals. Initial experimental results on the partially completed database are also presented, which demonstrated time-varying effect evidently.

1. Introduction

Speaker recognition is one kind of biometric authentication technology that can be used to automatically recognize a speaker's identity by using speaker-specific information contained in speech waves. Closely analogous to fingerprint, the term of voiceprint was created to stand for this speaker-specific information by pioneer researchers. Although they believed identifiable uniqueness did exist in each voice just as that of fingerprints, several questions were put forward at the same time, one of which was: "Does the voice of an adult change significantly with time? If so, how?" [1]. Obviously, researchers have been aware of the time-varying issue in speaker recognition from the very beginning.

In 1997, Sadaoki Furui summarized advances in automatic speaker recognition in decades and also raised an open question about the way to deal with long-term variability in people's voice [2]. It was suspected whether there was any systematic long-term variation that helped update speaker models to cope with the gradual changes. A similar idea was expressed in [3], where the authors argued that a big challenge to uniquely characterize a person's voice was that voice changes over time, either in the short-term (at different times of day), the medium-term (times of the year), or in the long-term (with age).

Furthermore, several researchers have observed performance degradation in presence of time intervals in practical systems. F. Soong *et al*. [4] concluded from experiments that the longer the separation between the training and the testing recordings, the worse the performance. Kato and Shimizu [5] also reported a significant loss in accuracy between two sessions separated by 3 months and aging was considered to be the cause [6].

Although performance degradation with time varying is a generally acknowledged phenomenon and it is widely assumed that speaker models should be updated from time to time to maintain representativeness, few researchers have figured out reasons behind this phenomenon exactly. No doubt a proper longitudinal voiceprint database is essential for this study.

The MARP corpus [7], including 21 sessions of free-flowing conversations over a three-year period of time for each speaker, has been the only one published so far. Each conversation was approximately ten minutes in length with an isolated partner who remained constant throughout the three years, and speakers were given suggested conversation topics. In the following study of aging effect [8], which largely focused on 32 speakers and 672 sessions from June 2005 to March 2008, it was pointed out that, while the

impact on speaker recognition accuracy between any two sessions was considerable, the long-term trend was statistically quite small, and it was argued that the detrimental impact was clearly not a function of aging or of the voice changing within this timeframe. However, the fact should not be neglected that in free-flowing conversations, speech contents are not fixed and a speaker's emotion, speaking style, or engagement can be easily influenced by his/her partner. Perhaps the aging effect is somewhat weaker than those evident ones and thus covered underneath. Hence, creation of a voiceprint database which specially focuses on the time-varying effect in speaker recognition is imperative for both research and practical applications.

In this paper, an on-going voiceprint database creation project is presented, aiming to examine solely the time-varying impact on speaker recognition and explore hidden factors causing possible performance degradation.

The paper is organized as follows. In Section 2 database design principles are presented and in Section 3 the database creation procedure is detailed. Initial experimental results showing performance degradation on the partially completed database are presented in Section 4 and conclusions are given in Section 5.

2. Database design principles

The time-varying effect is the only focus of this database creation project, therefore other factors such as recording equipments, software, conditions and environment should be kept as constant as possible throughout all recording sessions. Obviously, different from the MARP [7] corpus, we designed to create fixed-text read speech corpus. In the database design, two major factors were well considered, the prompt texts design and the time intervals design. Corresponding principles are listed as follows.

2.1. Fixed prompt texts

In order to focus on only the time-varying factor, other free variables during recording should be as few as possible. Therefore, speakers were requested to utter in a reading way with fixed prompt texts instead of free-style conversations. Prompt texts were designed to remain unchanged throughout all recording sessions. This is to avoid or at least reduce the impact of speech contents on speaker recognition accuracy. Furthermore, the general principle to retain well-balanced acoustic phenomena was also considered.

Prompt texts were organized in form of sentences and isolated words. Sentence utterances can be used to evaluate the performance of speaker recognition systems, while isolated words can be used to find variability in speech characteristics across sessions.

2.2. Gradient time intervals

In the MARP corpus, 21 sessions were recorded over a period of 33 months, but there was no information respecting how the organizers determined time intervals between adjacent sessions, and whether or not these sessions were of an approximately equal time interval.

Since there exists no precedent reference of time-interval design and it is costly and perhaps unnecessary to record in a fixed-length time interval for more than ten times to obtain a possible trend, gradient time intervals were used in this project. To be concrete, the first four sessions were recorded in an interval of approximately one week, the following four sessions in an interval of approximately one month, and so on, with the length of intervals increasing gradually, based on an assumption that the speaker recognition performance degrades drastically in the beginning, and not so much when the time difference between the testing and the training gets bigger. No matter the assumption is correct or not, however, the impacts of different time intervals can be easily analyzed in later experiments. Therefore, initial sessions can be of shorter time intervals, while following sessions of longer and longer time intervals.

3. Database creation

The details of the database creation procedure are given in this section. The prompt texts, recording sessions designing, speakers and recording conditions are described as follows.

3.1. Prompt texts

The prompt texts for the target database were made up of 100 Chinese sentences and 10 isolated Chinese words. Each speaker read the same prompt texts in each session. The length of each sentence ranges from 8 to 30 Chinese characters with an average of 15. Each isolated Chinese word contained 2 to 5 Chinese characters and was read five times in each session. Of the 10 isolated words, 5 were unchanged throughout all sessions just like the sentences, while the other 5 changed from session to session and reserved for future research of other purpose.

Chinese is a syllabic language with an Initial/Final structure where there are 21 Initials and 38 Finals [9]. Pronunciations of these Initials and Finals are strongly

influenced by their context, so the di-IF modeling was used and there are 1,523 different di-IFs in total [10]:

Initials -- *b*, *d*, *g*, *p*, *t*, *k*, *z*, *zh*, *j*, *c*, *ch*, *q*, *f*, *s*, *sh*, *x*, *h*, *m*, *n*, *l*, *r*, *y*, *w*;

Finals -- *a*, *o*, *e*, *i*, *i*1, *i*2, *u*, *v*, *er*, *ai*, *ei*, *ao*, *ou*, *ia*, *ie*, *ua*, *uo*, *ve*, *iao*, *iou*, *uai*, *uei*, *an*, *ian*, *uan*, *van*, *en*, *in*, *un*, *vn*, *ang*, *iang*, *uang*, *eng*, *ueng*, *ing*, *ong*, *iong*;

Di-IFs -- *f* + *a*, *b* + *u*, *ei* + *j*, ….

The acoustic coverage of the designed prompt texts is listed in Table 1.

Table 1. Acoustic coverage of prompt texts

	Number covered in prompt texts	*Total number*	*Percentage* (%)
Initials	23	23	100
Finals	38	38	100
di-IFs	1,183	1,523	78

3.2. Time intervals

This project was started in January 2010 and will be finished in 2012, where each speaker was/will be requested to utter 16 sessions. Five different time intervals are used: one week, one month, two months, four months and half a year, as illustrated in Figure 1. Suppose the j^{th} recording session be finished on day d_j as planned and the interval between the $(j\text{-}1)^{th}$ and j^{th} sessions be denoted by i_j. Considering that in actual recording it is unrealistic to make all speakers to record exactly on one specific day d_j, the session day d_j is made flexible to interval [d_j-0.25 i_j, d_j+0.25 i_j].

3.3. Speakers

60 fresh students were recruited for this three-year project when it was started, with 30 males and 30 females. This design of time intervals exactly does not make the summer and winter vacations the recording days when speakers would possibly be not on campus. Speakers were born in years between 1989 and 1993 with a majority in year 1990, and are from various departments, such as departments of computer science, biology, English, humanities, and journalism. Although they are from different parts of China, all of them speak standard Chinese well.

3.4. Recording conditions

An ordinary room in the laboratory is used for recording, where there is no burst noise but ambient noise in a low level. In the first session, all speakers were told what to do and what not to do in the recording room. They were requested to read the prompt texts in their normal speaking rate, while their volume can be controlled by the recording software. Most of speakers could complete a session in about 25 minutes smoothly.

Speech signals are digitalized at 8 kHz / 16 kHz sampling rates simultaneously in 16-bit precision. 8 recording sessions had been finished when this paper was submitted.

4. Database evaluation

A series of speaker verification experiments were performed on speech data from the first 7 recording sessions to evaluate the time-varying effect.

4.1. Experimental setup

The state-of-the-art 1024-mixture GMM-UBM (Gaussian Mixture Model - Universal Background Model) system was adopted, where 16-dimensional MFCCs (Mel-Frequency Cepstrum Coefficients) and their first derivatives were used as acoustic features.

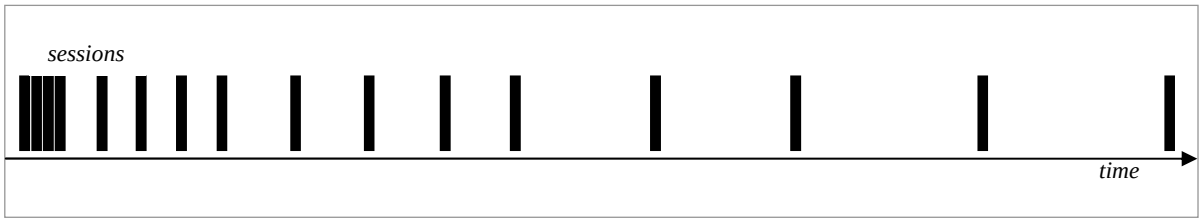

Figure 1. Illustration of different time intervals and session days

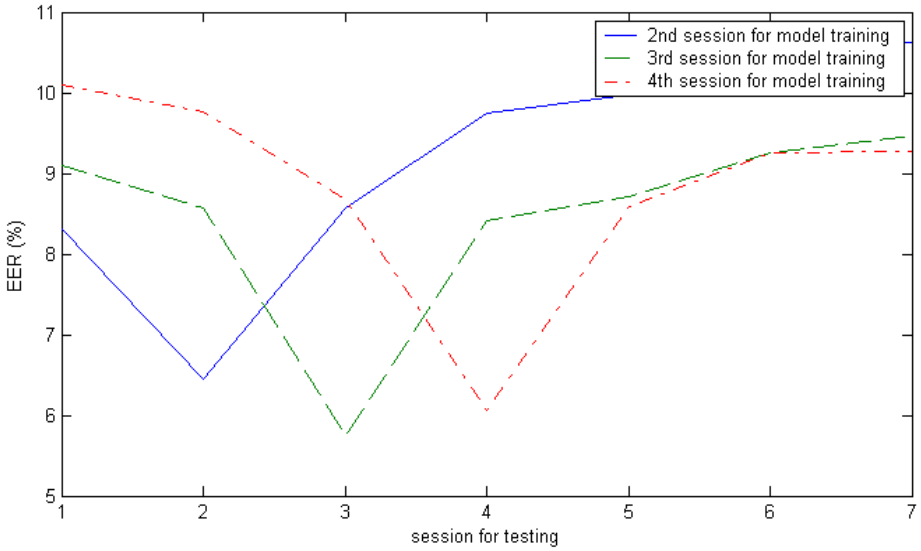

Figure 2. EER curves when using different sessions for model training

Each speaker model was trained using 3 sentences randomly selected from the entire 100 sentences with a length of about 10 seconds and all other sentences were used for testing, with each sentence ranging from 2 to 5 seconds.

Accordingly, there are 360,000 trials in each session.

4.2. Experimental results

Experimental results are shown in Figure 2, from which it can be seen that the speaker verification system performs best when training and testing utterances are from the same session, i.e., on the same recording dates. However, the performance gets worse and worse with the recording date difference between training and testing gets bigger.

5. Conclusions

In this paper, an on-going voiceprint database creation project is presented, aiming to examine solely the time-varying impact on speaker recognition and explore hidden factors causing possible performance degradation. In this speech database, speakers are requested to utter in a reading way with fixed prompt texts instead of free-style conversations throughout 16 sessions in a period of about three years. Sessions are of gradient time intervals where initial ones are of shorter time intervals and following ones of longer and longer time intervals.

Initial experimental results clearly demonstrated the time-varying effect, while the relationship between performance degradation and time intervals needs to be analyzed in the following research. Furthermore, finding out hidden factors behind this phenomenon is the ultimate goal of this time-varying research.

6. Acknowledgements

The authors would like to give their sincere thanks to members of voiceprint group in the Center. Without their pertinent suggestions and kind support, this database creation project will not function well.

7. References

[1] L.G. Kersta, “Voiceprint Recognition”, *Nature*, No. 4861, pp. 1253-1257, December 1962.

[2] S. Furui, “Recent Advances in Speaker Recognition”, *Pattern Recognition Letters*, Vol. 18, Iss. 9, pp. 859-872, September 1997.

[3] J. Bonastre, F. Bimbot, L. Boe, *et al*., “Person Authentication by Voice: A Need for Caution”, *Proc. of Eurospeech 2003*, pp. 33-36, Geneva, 2003.

[4] F. Soong, A. E. Rosenberg, L. R. Rabiner, *et al*., “A Vector Quantization Approach to Speaker Recognition”, *Proc. of ICASSP 1985*, Vol.10, pp. 387-390, Florida, 1985.

[5] T. Kato, and T. Shimizu, “Improved Speaker Verification over the Cellular Phone Network Using Phoneme-Balanced and Digit-Sequence Preserving Connected Digit Patterns”, *Proc. of ICASSP 2003*, Hong Kong, 2003.

[6] M. Hebert, “Text-Dependent Speaker Recognition”, *Springer Handbook of Speech Processing*, Springer-Verlag: Berlin, 2008.

[7] A. D. Lawson, A. R. Stauffer, E. J. Cupples, *et al*., “The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design and Initial Findings”, *Proc. of Interspeech 2009*, pp. 1811-1814, Brighton, 2009.

[8] A. D. Lawson, A. R. Stauffer, E. J. Cupples, *et al*., “Long Term Examination of Intra-Session and Inter-Session Speaker Variability”, *Proc. of Interspeech 2009*, pp. 2899-2902, Brighton, 2009.

[9] Jiyong Zhang, Fang Zheng, Jing Li, *et al*., “Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition”, *Proc. of Eurospeech 2001*, pp. 1617-1620, Aalborg, 2001.

[10] Zhenyu Xiong, Fang Zheng, Jing Li, and Wenhu Wu, “Au Automatic Prompting Texts Selecting Algorithm for di-Ifs Balanced Speech Corpus”, *Proc. of NCMMSC 7*, pp. 252-256, Xiamen, 2003.